Infodemiology and Infoveillance

Tracking Online Health Information and Cyberbehavior for Public Health

Gunther Eysenbach, MD, MPH

Introduction

Infodemiology, an emerging area of research at the crossroads of consumer health informatics and public health informatics, as well as infometrics and web analytics tools, can be defined as the science of distribution and determinants of information in an electronic medium, specifically the Internet, with the ultimate aim to inform public health and public policy. Infodemiology data (derived from unstructured, textual, openly accessible information produced and consumed by the public on the Internet, such as blogs, websites, and query and navigation data) can be collected and analyzed in near real-time. We developed a proof-of-concept infoveillance system called Infovigil, which can identify, archive, and analyze health-related information from Twitter and other information streams from Internet and social media sources. The system was developed to demonstrate and explore the potential of infoveillance for measuring public attention, attitudes, behavior, knowledge, and information consumption, as well as for syndromic surveillance, health communication, and knowledge translation research.

What Are Infodemiology and Infoveillance?

Imagine being a public health official, eHealth researcher, or behavioral scientist, and being able to look at a dashboard telling you in real-time what people are doing or feeling, much as economists can look at the Dow Jones stock index as a real-time measure of "what people are doing" (buying or selling), or at the VIX (volatility index, also called "fear index"), which provides metrics for implied feelings or attitudes, such as investor nervousness or fear.

The vision explored in this paper is to provide real-time metrics (presented as graphics, indices, and maps) of public behavior, opinion, knowledge, and attitudes to public health officials and policymakers, based on textual data harvested from the Internet. The amount of user-generated data on social media and other Internet-based venues "has made measurable what was previously immeasurable,"[1] and opens up a new intriguing area of research and development: the possibility to systematically mine, aggregate, and analyze these textual, unstructured data, to inform public health and public policy.

This emerging field has been called **infodemiology**,[1–3] originally in the context of identifying and monitoring misinformation,[2] and later in a study that showed that Google searches predicted influenza outbreaks[1] (later popularized by the Google Flutrends application).[4] Another term—**infoveillance**[3]— has been used for applications where infodemiology methods are employed for surveillance purposes.

The underlying idea of this emerging field is to measure the pulse of public opinion, attention, behavior, knowledge, and attitudes by tracking what people do and write on the Internet. The term **technosocial predictive analysis** also has been proposed,[5] although "prediction" does not capture the full range of possible applications of infodemiology. As will be discussed in more detail below, infodemiology goes beyond forecasting (prediction) and includes "nowcasting" (providing data for situational awareness on what the public does, knows, or feels about certain issues).

The term infodemiology is now widely used by others: in January 2011, there were almost 5000 Google hits for the term (as an aside, it should be noted that reporting and tracking the number of hits on Google for an emerging concept is a little infodemiology study in itself: The emergence of a new concept or term can be visualized by plotting the number of occurrences on the web as a knowledge translation or diffusion metric. We have done a similar analysis to illustrate the uptake of the new, recommended term H1N1 versus the popularly used "swine flu" during the 2009 H1N1 pandemic).[6]

Perhaps the term infodemiology is preferred because it intuitively conveys a few key concepts, including the fact

---

From the Centre for Global eHealth Innovation, Consumer Health & Public Health Informatics Lab, University Health Network, and the Department of Health Policy, Management, and Evaluation, University of Toronto, Toronto, Canada

Address correspondence to: Gunther Eysenbach, MD, MPH, Centre for Global eHealth Innovation, Consumer & Public Health Informatics Lab, University Health Network, 190 Elizabeth Street, Toronto M5G 2C4 Canada. E-mail: geysenba@gmail.com.

0749-3797/$17.00
doi: 10.1016/j.amepre.2011.02.006

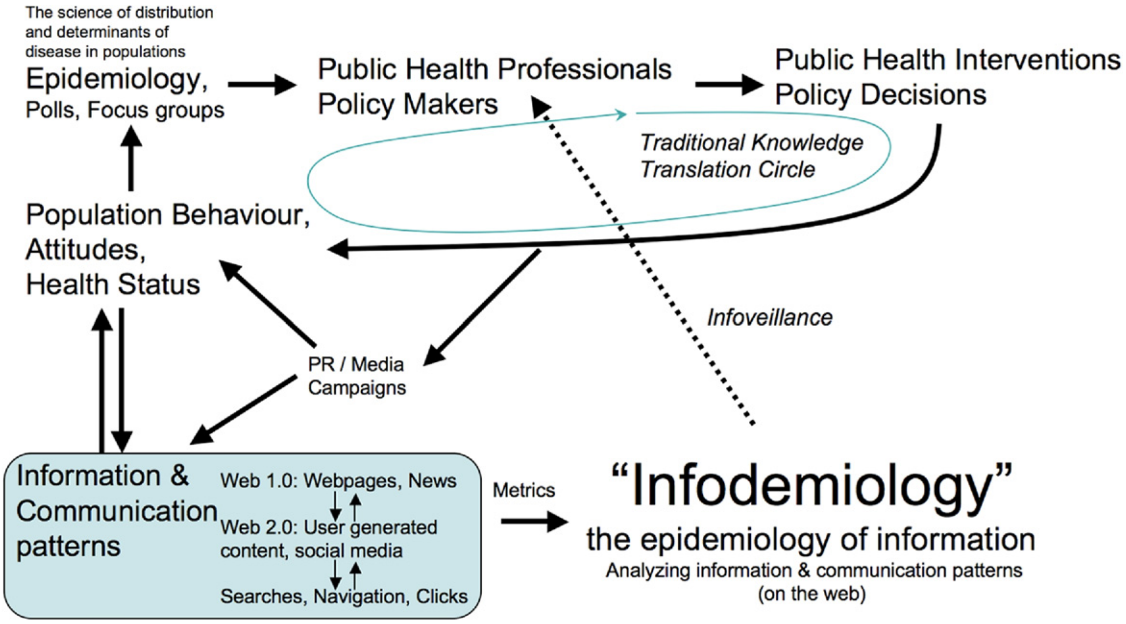

The science of distribution and determinants of disease in populations
Epidemiology, Polls, Focus groups
Public Health Professionals Policy Makers
Public Health Interventions Policy Decisions
Traditional Knowledge Translation Circle
Population Behaviour, Attitudes, Health Status
Infoveillance
PR / Media Campaigns
Information & Communication patterns
Web 1.0: Webpages, News
Web 2.0: User generated content, social media
Searches, Navigation, Clicks
Metrics
"Infodemiology"
the epidemiology of information
Analyzing information & communication patterns (on the web)

Figure 1. The role of infodemiology in public health

that (1) epidemiologic methods and terminology (e.g., prevalence) can be used to study and describe information in an information universe; (2) infodemiology provides data for public health decision making (again similar to the role of epidemiology); and (3) infodemiology takes a population-perspective (implied by the word *demos,* from Greek for *people*) on three different levels. First, people are the generators of this information; second, information has an effect on other people; and finally, it also reminds us of the fact that we have to look at populations of information units (e.g., many different websites) rather than an individual unit (e.g., web analytics of a single website) in order to obtain meaningful and robust data.

A more formal definition of infodemiology is "the science of distribution and determinants of information in an electronic medium, specifically the Internet, or in a population, with the ultimate aim to inform public health and public policy."[3] Infoveillance is "the longitudinal tracking of infodemiology metrics for surveillance and trend analysis."[3] With *information* we mean primarily unstructured, textual, openly accessible information produced and consumed by the public on the Internet. This can include, for example, search or navigation data (information demand), or postings (information supply) on websites, blogs, microblogs (Twitter), discussion boards, or social media. It can also include data on what people browse, buy, and read on the Internet, or social networking data (who we befriend or interact with) harvested from sites such as Facebook. Some of these data may actually be accessible in a structured format, but usually infodemiology implies some sort of free-text analysis. While the Internet is currently the main source of such information, in principle any consumer health informatics application (including, for example, personal health records, or even domotics applications such as intelligent kitchen appliances) and social media may produce data that may be harnessed for infodemiology and infoveillance approaches.

Figure 1 further illustrates the relationship between epidemiology (the science of the determinants and distribution disease) and infodemiology (the science of the determinants and distribution of information). The top cycle shows how traditional epidemiologic methods (e.g., surveys, clinical data) and data from polls or focus groups inform public health professionals and policymakers and affect public health interventions and policy decisions, which—augmented by the media and public relations campaigns—then (hopefully) have an impact on population behavior, attitudes, and ultimately health status. These outcomes are in turn picked up by traditional epidemiologic research methods, which again inform public health professionals and policymakers, and so on. This cycle is often a time-consuming process; for example, it may take months or years to determine whether a healthy-eating campaign has been successful in terms of affecting behavior, attitudes, knowledge, or even clinical outcomes. However, in the age of the Internet and social media, changes in population behavior, public attitudes, public attention, or health status are often reflected in immediate changes in information and communication patterns on the Internet.[1,3] If these changes could be picked up by infodemiology metrics, these data points could give some additional information to decision makers (we stress that infodemiology metrics cannot replace, but rather complement, traditional methods).

For example, changes in the health status of a population (more people getting influenza) leads to an increased search for influenza-related websites, more clicks on advertisements for influenza websites, more tweets and status updates on Facebook saying "I've got a cold," more orders in Internet pharmacies, possibly more book sales of influenza-related books, and so on. These infodemiology data will obviously be confounded and influenced by media reports (which in itself can be tracked and picked up by infodemiology methods). An "epidemic of fear" may exhibit similar characteristics as a true epidemic, so a triangulation taking into account, for example, traditional surveillance and epidemiologic methods as well as news reports in addition to monitoring user-generated data is required to determine the best course of action for public health if spikes and sudden changes in information patterns or chatter occur.

Infodemiology should not be misunderstood as having only practical applications in the context of infectious diseases (much as epidemiology is not only the science of epidemics). On the contrary, monitoring and combating behavioral risk factors for chronic diseases are other important application areas. Public health agencies could monitor, for example, the effectiveness or reach of a smoking-cessation campaign by tracking references to their campaign in blogs or discussion forums, or by monitoring whether their campaign has an impact on Twitter status updates along the lines of "I am trying to stop smoking." To some degree, many may already do this (using web analytics packages or media consultants to provide reports), but again, the true potential of infodemiology is unfolded only if data are embedded in a richer system that aggregates data from different sources and about different campaigns, topics, or issues.

Note that the arrow between population health status/attitudes/behavior and information patterns on the Internet in Figure 1 is bidirectional. That is to say that in some situations, a change in health status, attitudes, behavior, or knowledge leads to (or mirrors) changes in information and communication patterns, for example, in the context of an infectious disease outbreak. But conversely, changes in information and communication patterns (for example, through a media campaign or an anti-vaccination campaign) will have an impact on behavior and attitudes.

Examples of Infodemiology Applications

Examples for infodemiology applications include:

- the analysis of queries from Internet search engines to predict disease outbreaks (e.g., influenza)[1,4];
- mining status updates on microblogs such as Twitter for syndromic surveillance and situational awareness during a pandemic[6];
- identifying and monitoring of public health–relevant publications on the Internet (e.g., anti-vaccination sites, but also news articles or expert-curated outbreak reports)[2];
- detecting and quantifying disparities in health information availability[3];
- tracking the effectiveness of health marketing campaigns;
- extracting user-generated health outcomes data (e.g., from sites such as PatientsLikeMe) to construct patient-centered research instruments or to monitor drug side effects, off-label uses, or other medically interesting data[7–9];
- automated tools to measure information diffusion and knowledge translation (e.g., quantifying and visualizing the occurrence of certain terms and concepts over time).[6]

The previous examples of **H1N1** versus **swine flu** term occurrences belong in this last category, and so are tools such as article-level metrics at PLoS or the *Journal of Medical Internet Research* (JMIR). For example, JMIR monitors and counts mentionings of JMIR articles on Twitter (displayed as "top articles"), as a diffusion and impact metric (a Tweets Influence Index is also presented, which takes into account the number of followers for the respective tweets). This has overlaps with the field of scientometrics, with the difference that these metrics do not measure uptake within the scientific community (like citations), but by the general public.

Analyzing how people search and navigate the Internet for health-related information, as well as how they communicate and share this information, can provide valuable insights into the health-related behavior of populations. Another angle with which to look at infodemiology is the public engagement angle; closing the feedback loop between what messages are being sent out by public health agencies and how they resonate with the public can help agencies to tailor and target future communication and education strategies.

Advantages and Limitations

Infodemiology adds a novel set of methods to the toolbox of researchers and practitioners in the field of public health and policy research. The primary advantages are that— once set up—metrics are available in real time, can be collected automatically and inexpensively, and provide both quantitative and qualitative data. For example, we can do an initial qualitative analysis of tweets to set up a classification system to categorize tweets, and then do an automated analysis to quantitatively monitor the tweets in these categories prospectively.[1] Or, conversely, if we see unusual activity in our automated tracking, we can dig deeper (e.g., inspect the tweets leading to spikes) and explore why certain spikes occur.[6]

It is important to note that infodemiology has (just as any other method) certain limitations. For one, textual data can be messy and difficult to classify or interpret. Short keywords (such as Google search keywords) are easier to classify automatically but are harder to interpret semantically, as it is not clear why, for example, a person searching for *flu* is entering that keyword (does he have the flu? Is he writing a term paper about the flu? Has he read an article about the flu?). Longer blogs or websites are very information rich, but the semantics (meaning) are harder to extract automatically. We have focused our recent work on shorter status messages such as tweets because they are in a "sweet spot": long enough to provide depth and meaning and concise enough to facilitate rapid qualitative analysis or automatic classification. But tweets have other problems such as the use of abbreviations and texting jargons.

Representativeness is another limitation to keep in mind. Obviously, populations using the Internet (or the subpopulation of Internet users using social media) are not representative for the entire population (they are younger, more educated, have higher incomes, and are more likely reside in urban areas).[10] Just as with any other data collected for public health purposes, the potential biases need to be accounted for. Monitoring tweets during the H1N1 epidemic made sense because younger people were the ones primarily affected, but, for example, mining the attitudes and behaviors of older adults with Alzheimer disease would not be a very suitable infodemiology project.

Coding the geographic origin of information, or what geographic area it refers to, is sometimes another problem, depending on the data source. Data originating from social media can sometimes be linked back to the profile of the user, which may or may not contain geographic information, and which may or may not be the actual location of the user. Other types of infodemiology data sources (e.g., searches or webpage navigation patterns) may contain IP addresses, which—with some limitations—allow geocoding on a city level. Tweets sent from mobile devices may be geocoded, allowing the exact location to be determined.

Finally, depending on the level of analysis, novel privacy issues arise. For example, in the context of analyzing search queries, it has been shown that a re-identification of users from such data is possible, if the individual searches are linked by identifiers.[11] Automatic tracking and aggregation from published textual data sources (including the social web such as the public Twitter stream) does not normally raise privacy issues or even require ethics board approval (much as a systematic review or analysis of newspaper clippings would not require ethics review), but in some digital venues people may have a reasonable expectation of privacy. For example, an in-depth qualitative analysis of virtual communities creates issues that are similar to those arising in the context of analyzing discussion groups.[12] Tracking individual trajectories of users may create even bigger privacy concerns, in particular when they are not conducted in an automatic fashion and if personally identifiable information remains in the data set.

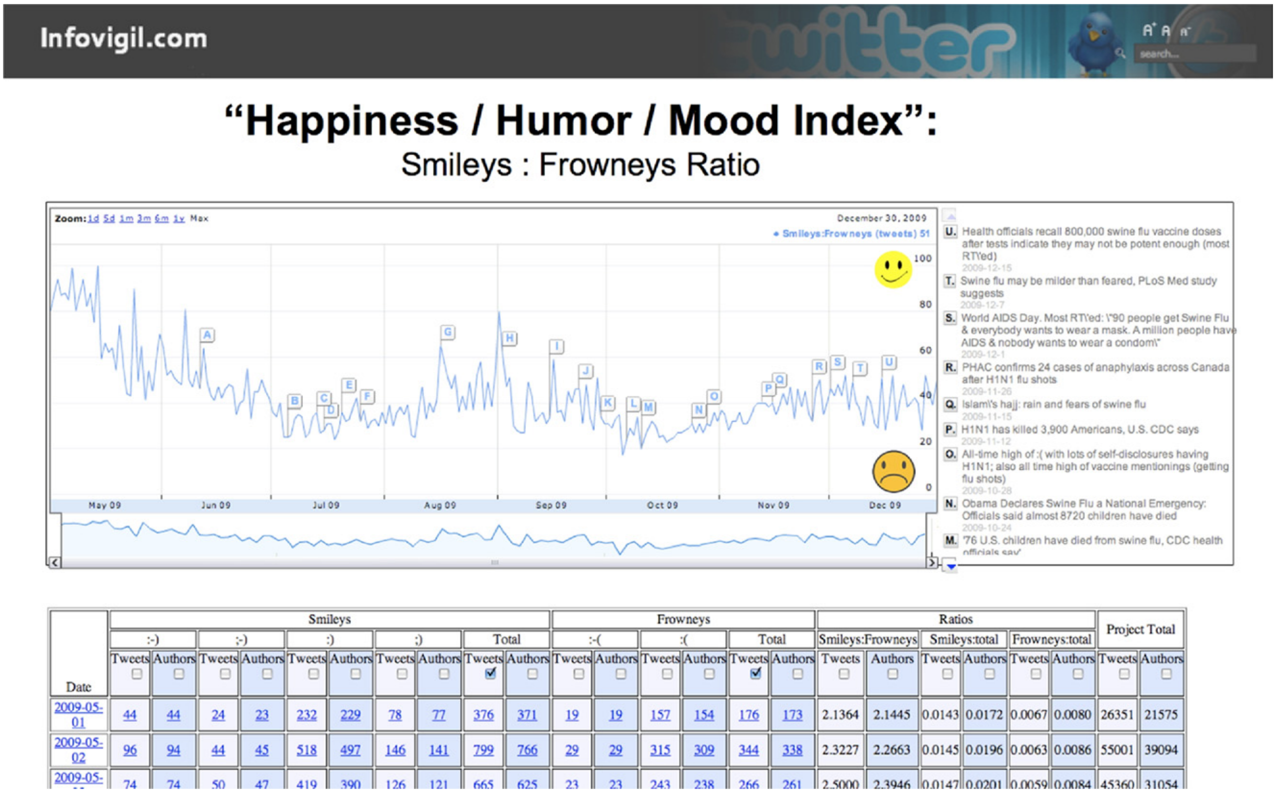
Infovigil

An open-source infodemiology system (dubbed Infovigil) is currently being developed at the Consumer Health & Public Health Informatics Lab in Toronto,[3,6] with the vision to provide a toolkit and "dashboard" for researchers and public health officials, and to conduct infoveillance projects by collecting, analyzing, and visualizing data from various sources on the Internet in real time. While we currently focus on Twitter streams, various other data sources can be plugged into the system (e.g., Google hits). Various indexes and indicators can be constructed, which show real-time sentiment, public opinion, public health–relevant behavior, inequities and disparities in the availability of health information or eHealth services, and other population-health and health policy–relevant metrics. Figure 2 provides a screenshot from the analysis page of the system. The researcher can define concepts (such as "smiley" and "frowny") and within each concept can define search keywords or patterns. The system plots ratios of concepts or prevalence rates of concepts within the data stream. Future iterations of the system will support natural language processing and maps.

Conclusion and Advice for Funders of Consumer Health Informatics Applications

Infodemiology is an emerging field, and we hope that policymakers, reviewers, and granting agencies recognize the potential of supporting research in this area and do not leave the field to proprietary players (such as Google Trends and Google Flutrends, who use closed "black box" algorithms and data).

Apart from this, both developers and sponsors of consumer health informatics applications have a role to play in implementing (or, in the case of funders, potentially even mandating) interfaces for infodemiology tools, so that relevant data can be aggregated across applications (we call these **infodemiology application programming interfaces** or ID-APIs). There is a growing appreciation



U. Health officials recall 800,000 swine flu vaccine doses after tests indicate they may not be potent enough (most RT\'ed)
2009-12-15

T. Swine flu may be milder than feared, PLoS Med study suggests
2009-12-7

S. World AIDS Day. Most RT\'ed: \"90 people get Swine Flu & everybody wants to wear a mask. A million people have AIDS & nobody wants to wear a condom\"
2009-12-1

R. PHAC confirms 24 cases of anaphylaxis across Canada after H1N1 flu shots
2009-11-26

Q. Islam\'s hajj: rain and fears of swine flu
2009-11-15

P. H1N1 has killed 3,900 Americans, U.S. CDC says
2009-11-12

O. All-time high of :(with lots of self-disclosures having H1N1; also all time high of vaccine mentionings (getting flu shots)
2009-10-28

N. Obama Declares Swine Flu a National Emergency: Officials said almost 8720 children have died
2009-10-24

M. '76 U.S. children have died from swine flu, CDC health officials say'

Date	Smileys										Frowneys						Ratios						Project Total	
	:-)		;-)		:)		;)		Total		:-(		:(		Total		Smileys:Frowneys		Smileys:total		Frowneys:total			
	Tweets ☐	Authors ☐	Tweets ☐	Authors ☐	Tweets ☐	Authors ☐	Tweets ☐	Authors ☐	Tweets ☑	Authors ☐	Tweets ☐	Authors ☐	Tweets ☐	Authors ☐	Tweets ☑	Authors ☐	Tweets ☐	Authors ☐	Tweets ☐	Authors ☐	Tweets ☐	Authors ☐	Tweets ☐	Authors ☐
2009-05-01	44	44	24	23	232	229	78	77	376	371	19	19	157	154	176	173	2.1364	2.1445	0.0143	0.0172	0.0067	0.0080	26351	21575
2009-05-02	96	94	44	45	518	497	146	141	799	766	29	29	315	309	344	338	2.3227	2.2663	0.0145	0.0196	0.0063	0.0086	55001	39094
2009-05-	74	74	50	47	419	390	126	121	665	625	23	23	243	238	266	261	2.5000	2.3946	0.0147	0.0201	0.0059	0.0084	45360	31054

Figure 2. Infovigil screenshot of real-time emoticon mood analysis of Tweets related to H1N1

among funders of health research that research results and data from research funded by public money should be openly accessible and reusable,[13] and we think that this should extend to (anonymized) usage data produced in real time by publicly funded consumer health applications. We suggest a concerted effort to define such open standards, interfaces, policies, and systems, to enable mining and analyzing these data on a large scale.

Publication of this article was supported by the National Institutes of Health.

No financial disclosures were reported by the author of this paper.

References

1. Eysenbach G. Infodemiology: tracking flu-related searches on the web for syndromic surveillance. AMIA Annu Symp Proc 2006:244–8.
2. Eysenbach G. Infodemiology: the epidemiology of (mis)information. Am J Med 2002;113(9):763–5.
3. Eysenbach G. Infodemiology and infoveillance: framework for an emerging set of public health informatics methods to analyze search, communication and publication behavior on the Internet. J Med Internet Res 2009;11(1):e11.
4. Ginsberg J, Mohebbi M, Patel R, Brammer L, Smolinski M, Brilliant L. Detecting influenza epidemics using search engine query data. Nature 2009;457(7232):1012–4.
5. Boulos K, Sanfilippo A, Corley C, Wheeler S. Social Web mining and exploitation for serious applications: Technosocial Predictive Analytics and related technologies for public health, environmental and national security surveillance. Comput Methods Programs Biomed 2010;100(1):16–23.
6. Chew C, Eysenbach G. Pandemics in the age of Twitter: content analysis of tweets during the 2009 H1N1 outbreak. PLoS ONE 2010;29(5):e14118.
7. Wicks P, Massagli M, Kulkarni A, Dastani H. Use of an online community to develop patient-reported outcome instruments: the Multiple Sclerosis Treatment Adherence Questionnaire (MS-TAQ). J Med Internet Res 2011;13(1):e12.
8. Wicks P, Massagli M, Frost J, Brownstein C, Okun S, Vaughan T. Sharing health data for better outcomes on PatientsLikeMe. J Med Internet Res 2010;12(2):e19.
9. Frost J, Okun S, Vaughan T, Heywood J, Wicks P. Patient-reported outcomes as a source of evidence in off-label prescribing: analysis of data from PatientsLikeMe. J Med Internet Res 2011;13(1):e6.
10. Statistics Canada. Canadian Internet Use Survey, 2009. www.webcitation.org/5w9gvNAWt.
11. CNET News. AOL's disturbing glimpse into users' lives. www.webcitation.org/5w9ev2kNY.
12. Eysenbach G, Till J. Ethical issues in qualitative research on internet communities. BMJ 2001;232:1103.
13. National Institutes of Health. Final NIH Statement on Sharing Research Data. www.webcitation.org/5w9XuU1Mw.